ABSTRACT

We propose a novel method for estimating speech rate based on the durations of similar patterns as a first step in determining the relation between various speaking styles used in everyday conversation and speaker intentions or attitudes. Whereas most methods of determining speaking rate require manually obtained label information or linguistic knowledge, the proposed method uses patterns of speech-sound sequences that occur relatively frequently in dialogue speech, as detected from the speech waveform information alone. For use as an index of paralinguistic information encoded in the various speaking styles used in everyday conversation, the method allows us to analyze changes in speech rate with respect to speaker intentions or speaker state. The method allows us to work with untranscribed speech, processing not only the speech patterns demarcated by pauses, but also the patterns embedded in a sentence utterance.

In addition, this method of pattern extraction processing is as a general preprocessor for speaking style analysis of large scale speech data, since it can provide a basic framework for extraction and analysis of any variable features of dialogue speech data, i.e., it is applicable not only to speech rate but also to various acoustic or prosodic features such as fundamental frequency, power, and voice quality, to enable an estimate of local settings normalised per speaker without access to a transcription or labels.

2. PATTERN EXTRACTION METHOD

Compared to the clean speech of laboratory or studio recordings, in dialogue speech, many utterances undergo extreme phonetic modification (e.g., prolongation of vowels or gemination of consonants, elision or deletion of sounds, etc.) and transcription is an expensive and time-consuming process, particularly if it is to be done at the phonetic level. Even if it is to be used as a word-level target for automated segmentation, the written transcription itself is not necessarily suitable as a specification of the speech sounds.

For example, although speech rate in Japanese is usually specified as a period of time divided by the number of mora (or syllables) it contains, when there is insertion or prolongation of vowels or gemination of consonants, for example, it becomes difficult to specify the number of mora.

Although application of speech recognition technology to dialogue speech data has improved considerably in recent years, many spontaneous speech utterances are not registered either in the lexical dictionary or the language model. Yet these non-verbal speech noises are common in dialogue data, and though they signal important affective information, they also serve as a cause of recognition errors.

It is of course possible to add such non-verbal speech items to the dictionary or language model from existing transcriptions, but we note that this type of dialogue speech is very dependent on speakers and speaker-hearer conditions, and there is no guarantee that we can obtain a sufficient lexical set only by increasing the amount of transcribed data.

On the other hand, if similar types of speech patterns can
be extracted or detected automatically from dialogue speech, the duration of these similar patterns can be directly compared, and we can make an estimate of speech rate change by sampling the patterns at (irregular) intervals throughout the speech.

Therefore, we propose a method for extracting similar speech patterns automatically, based on acoustic information, as a component technology for the analysis of speaking style and other prosodic features that slowly change throughout the course of an utterance.

2.1. Phonetic sequencing

Fig. 1 illustrates the method. First, speech waveform information is decoded to phoneme character sequences using a speech recognizer [7, 8] (based on the following acoustic features: 12 dimensional MFCC + 12 dimensional Δ MFCC + Δ log power) in the ‘phonetic typewriter’ mode of the speech recognition engine Julius [9]. An ergodic grammar which permits all 43 phonemes to connect freely to each other is used for the ‘recognition’. In this method, the phoneme character sequence is not considered as text information related to utterance content, but as a mere symbol sequence for carrying out an indexing of the time series patterns of the acoustic features. We do not require speech recognition accuracy based on a correspondence of the recognition result with a transcribed text.

Because previous research [10] has indicated that speaker adaptation in phonemic segmentation based on transcribed text can improve accuracy better than speaker-independent HMMs, we adapted a monophone HMM using a set of phonemically balanced sentences read by the same speaker who provided the dialogue speech corpus (see 3.1 below).

2.2. Pattern extraction

Next, all patterns which appeared 50 times or more in the recognition result were extracted using the Multigram Software Package [11] which was designed for statistical natural-language processing.

The patterns were determined by dividing each label-pair sequence into pattern sequences using the maximum-likelihood-based VI-train command of the Multigram Software. The configuration was as follows: up to 10 labels can be considered as forming a pattern (or multigram dictionary entry). Each pattern must occur at least 50 times in the corpus, and the number of iterations for the likelihood estimation and pattern dictionary construction is 10. This process results in a dictionary of label-sequences that commonly co-occur, and an estimate of the occurrence likelihood for each entry.

2.3. Pattern grouping

The start time and end time are assigned to each speech pattern by post-processing of the recognizer output, and the z-scores of these segment durations are calculated for all tokens in each pattern group.

3. EXPERIMENTS

In order to check the validity of the proposed method, an evaluation experiment was conducted using a subset of the same dialogue speech corpus. The validity was evaluated by comparing the patterns obtained by the proposed method (henceforth, ‘recognition patterns’) without the use of a transcription, and the equivalent patterns obtained from a phonemic segmentation based on the transcribed text (henceforth, ‘transcription patterns’) which we assume to be more accurate since they are more expensive to obtain. If the same result is obtained for recognition patterns and transcription patterns, it can be said that the validity of the proposed method is high.

3.1. Corpus

A section of the spontaneous dialogue speech corpus from the JST/CREST-ESP project [1, 12, 13, 14] was used as the speech material.

This corpus has the following features:

- Spontaneous dialogue speech over the telephone, recorded using high-quality head-mounted microphone direct to local disc at 44.1kHz.
- A young adult female speaker of Japanese
- Data recorded over period of 2 years (more than 250 hours of speech data)
- Each dialogue lasts between 6 and 30 minutes.
- Includes various inter-personal relationships (parents, husband, children, relatives, friends, others)

We used 145,152 utterances from 792 dialogues for the training. Each utterance was determined from a manual transcription and defined as a ‘minimal meaningful speech unit’. Processing was performed with the proposed method, based only on speech waveform information alone, and no transcribed text or other information relating to the utterance content was used except for evaluation.

3.2. Phonemic segmentation as evaluation criterion

For the evaluation, we used similar patterns determined from the phoneme sequence output by a speech recognition engine (Julius [9]) used in alignment mode and fed with a hand-made transcription of each utterance in place of a grammar module.

As mentioned above, we used the VI-train command of the Multigram Package [11] to provide the dictionary of label-sequence patterns and the unit sequences. The settings of HMMs for phonemic segmentation and configuration of pattern extraction were as described in section 2.

Because the manually transcribed text is in Japanese orthography (and therefore mixes Chinese kanji characters with the phonetic kana alphabet) we used the public-domain Japanese morphological analyzer ChaSen [15] to produce a sequence of phonemic characters representing the text as a basis for the automatic segmentation.

3.3. Preliminary experiment

The number of labels making up the multigram dictionary codebook varied in the range of 1 ≤ N ≤ 6. No patterns longer than 6 occurring more than 50 times were extracted.
We found that there were many more fragmentary patterns of 3 phonemes for the recognition patterns, than for the transcription patterns. Moreover, the patterns incorporating 3 labels or fewer correspond to very many speech segments and may be greatly influenced by neighboring phonemes, so the variation of these segments will be great when compared with others of the ‘same’ pattern. On the other hand, patterns of 5 labels correspond to a smaller number of speech segments of a specific type, and the influence of surrounding phoneme variations can be considered to be comparatively small, so we limited our evaluation to only the changes of speech rate observed within the 5-label patterns.

3.4. Evaluation measure

In this section, we verify the validity of the pattern extraction results by the proposed method and confirm the speech rate estimation (based on z-scores of pattern durations in the group distributions of each pattern) by comparison with the case of transcription patterns.

First, if there is any overlap between speech segments corresponding to the recognition pattern and segments corresponding to the transcription patterns, then we judge the recognition pattern to be sufficiently aligned or correct. Next, the validity of the proposed method is evaluated at these well-aligned points.

For the correspondingly aligned data, the agreement between the pattern duration z-scores from recognition patterns and those from transcription patterns was checked. To the extent that both z-scores correspond, the possibility that speech rate can be appropriately measured by the proposed method is confirmed.

Furthermore, correlation between the proposed method and a conventional speech rate which used transcription was confirmed for all extracted patterns, for there were some recognition patterns which had no transcription pattern counterparts due to mis-recognition.

4. RESULTS AND DISCUSSION

4.1. Reliability of speech segment extraction

Table 1 shows the ten most frequent patterns of multigram code-words of length 5. Many include silence symbols (/sp/ or /q/) from the recognizer, but we selected the interjection /naNka/ for further manual inspection.

<table>
<thead>
<tr>
<th>pattern</th>
<th>number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;sp&gt;ru:Nq</td>
<td>162</td>
</tr>
<tr>
<td>&lt;sp&gt;nu:Nq</td>
<td>161</td>
</tr>
<tr>
<td>&lt;sp&gt;uu:Nq</td>
<td>152</td>
</tr>
<tr>
<td>kraqte</td>
<td>134</td>
</tr>
<tr>
<td>tokoa</td>
<td>133</td>
</tr>
<tr>
<td>toqte</td>
<td>118</td>
</tr>
<tr>
<td>naNka</td>
<td>116</td>
</tr>
<tr>
<td>&lt;sp&gt;moNq</td>
<td>107</td>
</tr>
<tr>
<td>moqte</td>
<td>104</td>
</tr>
<tr>
<td>koqto</td>
<td>99</td>
</tr>
</tbody>
</table>

We found 116 speech segments corresponding to /naNka/ in both the transcription patterns and the recognition patterns. Of these, 81 were correctly aligned (Table 2).

4.2. Speech rate based on z-score of pattern duration

The larger the z-score of pattern duration, the slower the speaking rate of the corresponding speech segment can be assumed in the proposed method. Because both distributions of duration from the 116 recognition patterns and transcription patterns are similar to a log normal distribution, duration is computed as z-scores after being converted to logarithmic values. As a result, both distributions of duration patterns are close to normal distribution (Fig. 2)

Table 2: Result of listening test

<table>
<thead>
<tr>
<th>word including pattern</th>
<th>meaning in english</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>naNka</td>
<td>interjection</td>
<td>96</td>
</tr>
<tr>
<td>naNkai</td>
<td>how many times</td>
<td>16</td>
</tr>
<tr>
<td>naNka</td>
<td>question</td>
<td>2</td>
</tr>
<tr>
<td>naNka</td>
<td>etc.</td>
<td>1</td>
</tr>
<tr>
<td>naNka</td>
<td>something</td>
<td>1</td>
</tr>
<tr>
<td>total</td>
<td>–</td>
<td>116</td>
</tr>
</tbody>
</table>

As a result of the checks on the 81 correct data in the 116 segments automatically extracted by the proposed method, the duration z-scores appear equal to the duration z-score of the transcription patterns in general. So, it is thought appropriate to predict speech rate based on z-score of recognition patterns obtained by the proposed method (Fig. 3).
The correlation between the proposed method and vowel-per-second rate [16] is shown in Fig. 4.

![Figure 4: Correlation between proposed method and conventional method. The ordinate illustrates the z-scores of all extracted recognition patterns based on the proposed method. The abscissa illustrates vowel-per-second rate of the speech segments corresponding the recognition patterns.](image)

Although overall correlation is $R = -0.356$, there seems to be 3 groups of correlation. The correlation computed for each group is comparatively high and patterns in each group has characteristic structure, i.e., number of vowels included in the recognition pattern (Table 3).

Table 3: 3 groups of correlation

<table>
<thead>
<tr>
<th>group</th>
<th>correspondence formula</th>
<th>correlation</th>
<th>vowels</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>$y = -0.5x + 1.8$</td>
<td>-0.693</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>$y = -1.0x + 5.5$</td>
<td>-0.551</td>
<td>2</td>
</tr>
<tr>
<td>c</td>
<td>$y = -2.0x + 8.0$</td>
<td>-0.592</td>
<td>3</td>
</tr>
</tbody>
</table>

5. CONCLUSION

We proposed a method for analyzing speaking rate based on the duration distribution of the estimated speech patterns, after automatically extracting sets of speech patterns which are not dependent on text information and which appear 50 times or more in a dialogue using only audio information, as the first step towards analyzing the various speaking styles encountered in a large-scale dialogue speech corpus. By this method, speech rate analysis can be considered as almost equivalent to the case where a transcribed text is used. We will make the proposed method more precise by reconsidering the structure of extracted patterns.

The pattern extraction technique of the proposed method is effective as a preprocessing technique for analyzing speaking styles based on a large amount of dialogue speech data, and provides a fundamental framework for extracting and analyzing the various features of paralinguistic information in speech. It is also applicable to the analysis of fundamental frequency, power, voice quality, etc.

6. ACKNOWLEDGEMENTS

We would like to thank the JST/CREST-ESP project for providing the dialogue speech data. We thank Dr. Parham MOKHTARI and Dr. Carlos Toshinori ISHI of the JST/CREST-ESP project for useful advices, and Mr. Keisuke NAKAMURA at NAIST for guidance with the use of speaker adaptation in HMM recognition.

7. REFERENCES